The Distribution of the Bivariate Kaplan-Meier Estimate

By

Ronald C. Pruitt

Technical Report No. 517

School of Statistics

University of Minnesota

3 August 1988

Abstract

In this note, conditions under which the bivariate Kaplan-Meier estimate of Dabrowska (1988) is not a proper survival function are given. All points assigned negative mass are identified.

Dabrowska (1988) introduced a multivariate survival curve estimate. In her paper she points out that her estimate may fail to be monotone and hence not be a survival function. This paper describes under what conditions Dabrowska's bivariate estimate is not a survival function.

Throughout we follow the notation of Dabrowska (1988). We wish to infer about a bivariate distribution $\vec{T} = (T_1, T_2)$ subject to censoring. Assume $\vec{T}$ and the censoring variable $\vec{Z} = (Z_1, Z_2)$ are defined on a common probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and have survival functions $F(s,t) = \Pr(T_1 > s, T_2 > t)$ and $G(s,t) = \Pr(Z_1 > s, Z_2 > t)$. The observable random variables are given by $\vec{Y} = (Y_1, Y_2)$ and $\vec{\delta} = (\delta_1, \delta_2)$, where $Y_i = T_i \wedge Z_i$ and $\delta_i = 1[T_i = Y_i]$, for $i = 1, 2$. To estimate F, suppose we have a sample $(\vec{Y}_i, \vec{\delta}_i)$, $i = 1, \ldots, n$ which consists of independent, identically distributed copies of $(\vec{Y}, \vec{\delta})$. Let

$$\begin{aligned}
\hat{H}(s,t) &= n^{-1} \sum 1[Y_{1i} > s, Y_{2i} > t] \\
\hat{K}_1(s,t) &= n^{-1} \sum 1[Y_{1i} > s, Y_{2i} > t, \delta_{1i} = 1, \delta_{2i} = 1] \\
\hat{K}_2(s,t) &= n^{-1} \sum 1[Y_{1i} > s, Y_{2i} > t, \delta_{1i} = 1] \\
\hat{K}_3(s,t) &= n^{-1} \sum 1[Y_{1i} > s, Y_{2i} > t, \delta_{2i} = 1].
\end{aligned}$$

These functions can be used to estimate the bivariate cumulative hazard function by

$$\begin{aligned}
\hat{\Lambda}_{11}(s,t) &= \int_0^s \int_0^t \hat{K}_1(du, dv) / \hat{H}(u-, v-) \\
\hat{\Lambda}_{10}(s,t) &= -\int_0^s \hat{K}_2(du, t) / \hat{H}(u-, t) \\
\hat{\Lambda}_{01}(s,t) &= -\int_0^t \hat{K}_3(s, dv) / \hat{H}(s, v-).
\end{aligned}$$

With $f(\Delta x) = f(x) - f(x-)$, define

$$\hat{L}(\Delta u, \Delta v) = \frac{\hat{\Lambda}_{10}(\Delta u, v-)\hat{\Lambda}_{01}(u-, \Delta v) - \hat{\Lambda}_{11}(\Delta u, \Delta v)}{\{1 - \hat{\Lambda}_{10}(\Delta u, v-)\}\{1 - \hat{\Lambda}_{01}(u-, \Delta v)\}}, \tag{1}$$

if the denominator of the right hand side is non-zero, and otherwise let $\hat{L}(\Delta u, \Delta v) = 0$. Dabrowska's estimate is

$$\hat{F}(s,t) = \hat{F}(s,0)\hat{F}(0,t) \prod_{\substack{0<u\leq s \\ 0<v\leq t}} \{1 - \hat{L}(\Delta u, \Delta v)\},$$

where $\hat{F}(s,0)$ and $\hat{F}(0,t)$ are the marginal Kaplan-Meier estimates.

We can now state the theorem. To simplify the form of the result, restrict attention to the case of absolutely continuous $\vec{Y}$.

Theorem 1 *Assume the distribution of $\vec{Y}$ is absolutely continuous. With probability one, Dabrowska's estimate is a discrete measure and negative mass is assigned in accordance with Lemmas 3–6.*

Proof: Restrict attention to the case when $Y_{i1}, \ldots, Y_{in}$ are all distinct for $i = 1, 2$. First note that mass is concentrated on the set of points $S = \{(y_1, y_2) : y_1 = Y_{1i}, y_2 = Y_{2j}, \delta_{1i} = 1,$ and $\delta_{2j} = 1$ for some $1 \leq i, j \leq n\}$. This can be seen by observing that the mass assigned to a point (s,t) may be written

$$\begin{aligned}
\hat{F}(\Delta s, \Delta t) &= \prod_{\substack{0<u<s \\ 0<v<t}} \{1 - \hat{L}(\Delta u, \Delta v)\} \times \Big\{ [\hat{F}(s-,0) - \hat{F}(s,0) \prod_{0<v<t} \{1 - \hat{L}(\Delta s, \Delta v)\}] \\
&\qquad \times [\hat{F}(0,t-) - \hat{F}(0,t) \prod_{0<u<s} \{1 - \hat{L}(\Delta u, \Delta t)\}] \\
&\qquad - \hat{L}(\Delta s, \Delta t) \hat{F}(s,0) \hat{F}(0,t) \prod_{0<u<s} \{1 - \hat{L}(\Delta u, \Delta t)\} \prod_{0<v<t} \{1 - \hat{L}(\Delta s, \Delta v)\} \Big\} \\
&\stackrel{\text{def}}{=} R_0(s,t) \{R_1(s,t) R_2(s,t) - R_3(s,t)\}, \qquad (2)
\end{aligned}$$

and noting that the set of points where $\hat{L}(\Delta u, \Delta v)$ is non-zero is contained in S, and the marginal Kaplan-Meier estimates only both change value on points in S. There are seventeen possible cases, the case when $i = j$ and the sixteen cases indicated in Figure 1. By symmetry only plots in Figure 1 on or beneath the diagonal need to be considered, for example plots 4 and 13 only differ in the labelling of the variables. These fall into four cases which are covered by Lemmas 3–6. □

Before stating and proving Lemmas 3–6 we give an auxiliary lemma which contains the essential ideas. Let $n_{u,v} = n\hat{H}(u-, v-)$ be the number of observations in $[u, \infty) \times [v, \infty)$.

Lemma 2 *Fix k, with $1 \leq k \leq n$ and $\delta_{2k} = 1$. Let $s = Y_{1k}$ and $t = Y_{2k}$. The following implications hold:*

1. *If $\delta_{1k} = 0$, $R_2(x,t) > 0$ for any $x \geq 0$.*
2. *If $\delta_{1k} = 1$, one of the following hold:*

(a) If $x \leq s$, $R_2(x,t) > 0$.

(b) If $x > s$, one of the following hold:

i. If $n_{s,t} = 1$, $R_2(x,t) > 0$.

ii. If $n_{s,t} > 1$, $R_2(x,t) \leq 0$ with equality if and only if the set $C = \{(Y_{1m}, Y_{2m}) : Y_{1m} < s, Y_{2m} > t, \delta_{1m} = 0\}$ is empty.

Proof: Since the observed values are all distinct,

$$\hat{\Lambda}_{01}(u-, \Delta t) = \begin{cases} (n_{u,t})^{-1} & \text{for } u \leq s \\ 0 & \text{for } u > s \end{cases}.$$

Combining this with similar equations for $\hat{\Lambda}_{10}$ and $\hat{\Lambda}_{11}$,

$$1 - \hat{L}(\Delta u, \Delta t) = \begin{cases} \frac{n_{u,t}(n_{u,t}-2)}{(n_{u,t}-1)^2} & \text{if } u < s \text{ and } \hat{K}_2(\Delta u, t) < 0 \\ \frac{n_{u,t}}{(n_{u,t}-1)} & \text{if } u = s, \hat{K}_2(\Delta s, t) < 0, \text{ and } n_{s,t} > 1 \\ 1 & \text{if } u = s, \hat{K}_2(\Delta s, t) < 0, \text{ and } n_{s,t} = 1 \\ 1 & \text{if } u > s \text{ or } \hat{K}_2(\Delta u, t) = 0 \end{cases}.$$

Now 1, 2a, and 2(b)i follow since $R_2(x,t) \geq \hat{F}(0,t-) - \hat{F}(0,t) > 0$. It is also clear that when $x > s, \delta_{1k} = 1$, and $n_{s,t} > 1$,

$$\prod_{0<u<x} \{1 - \hat{L}(\Delta u, \Delta t)\} \geq \frac{n_{0+,t}}{(n_{0+,t} - 1)}, \tag{3}$$

with equality if and only if the set C of 2(b)ii is empty. This finishes the proof after noting that $\hat{F}(0,t-)/\hat{F}(0,t) = n_{0+,t}/(n_{0+,t} - 1)$. □

We can now describe when negative mass is assigned by Dabrowska's estimate.

Lemma 3 (Type I) *Assume $Y_{1i} < Y_{1j}, Y_{2i} > Y_{2j}, \delta_{1i} = 1$, and $\delta_{2j} = 1$. Then negative mass is assigned to the point (Y_{1i}, Y_{2j}) if and only if the set of points $D = D_1 \cup D_2$ is non-empty, where $D_1 = \{(Y_{1k}, Y_{2k}) : Y_{1k} < Y_{1i}, Y_{2k} > Y_{2j}$, and $\delta_{1k} = 0$ for some $k = 1, \ldots, n\}$ and $D_2 = \{(Y_{1k}, Y_{2k}) : Y_{1k} > Y_{1i}, Y_{2k} < Y_{2j}$, and $\delta_{2k} = 0$ for some $k = 1, \ldots, n\}$.*

Proof: This covers plots 1, 2, 5, and 6 of Figure 1. Let $s = Y_{1i}, t = Y_{2j}$, and note $R_0(s,t) > 0$. From (3),

$$\prod_{0<u<s} \{1 - \hat{L}(\Delta u, \Delta t)\} \geq \frac{n_{0+,t}}{(n_{0+,t}-1)} \frac{(n_{s,t}-1)}{n_{s,t}},$$

with equality if and only if D_1 is empty. Applying this to (2) gives

$$\begin{aligned} \frac{\hat{F}(\Delta s, \Delta t)}{R_0(s,t)} \leq\ & \hat{F}(s,0)\frac{n_{s,0+}}{(n_{s,0+}-1)}\frac{1}{n_{s,t}}\hat{F}(0,t)\frac{n_{0+,t}}{(n_{0+,t}-1)}\frac{1}{n_{s,t}} \\ & -\frac{1}{(n_{s,t}-1)^2}\hat{F}(s,0)\hat{F}(0,t)\frac{n_{s,0+}}{(n_{s,0+}-1)}\frac{(n_{s,t}-1)}{n_{s,t}}\frac{n_{0+,t}}{(n_{0+,t}-1)}\frac{(n_{s,t}-1)}{n_{s,t}} = 0, \end{aligned}$$

with equality if and only if D_1 and D_2 are both empty. □

Lemma 4 (Type II) *Assume $Y_{2i} < Y_{2j}$ and $\delta_{2j} = 1$. Also assume that if $\delta_{1j} = 1$, then $Y_{1i} < Y_{1j}$. Then negative mass is assigned to the point (Y_{1i}, Y_{2j}) if and only if the set $E_1 = \{(Y_{1k}, Y_{2k}) : Y_{1k} > Y_{1i}, Y_{2k} < Y_{2i}$, and $\delta_{2k} = 0$ for some $k = 1, \ldots, n\}$ is non-empty, the set $E_2 = \{(Y_{1k}, Y_{2k}) : Y_{1k} > Y_{1i}$ and $Y_{2k} > Y_{2i}$ for some $k = 1, \ldots, n\}$ is non-empty, and $R_0(Y_{1i}, Y_{2j}) > 0$. Also, if $Y_{1i} < Y_{1j}$, then $R_0(Y_{1i}, Y_{2j}) > 0$ and E_2 is non-empty.*

Proof: This covers plots 13, 14, and 15 of Figure 1. Plots 4, 8, and 12 are obtained by interchanging the roles of the variables. Note that $R_3(Y_{1i}, Y_{2j}) = 0$, and by Lemma 2 $R_2(Y_{1i}, Y_{2j}) > 0$, and $R_1(Y_{1i}, Y_{2j}) < 0$ if and only if E_1 and E_2 are each non-empty. This proves the first assertion. For the second assertion, $1 - \hat{L}(\Delta u, \Delta v) > 3/4$ for $0 \leq u < Y_{1i}$ and $0 \leq v < Y_{2j}$ when $Y_{1i} < Y_{1j}$. □

Lemma 5 (Type III) *Assume $Y_{1i} > Y_{1j}, Y_{2i} < Y_{2j}$, and $\delta_{1i} = \delta_{1j} = \delta_{2i} = \delta_{2j} = 1$. Then (Y_{1i}, Y_{2j}) is assigned negative mass if and only if [the set $F_1 = \{(Y_{1k}, Y_{2k}) : Y_{1k} > Y_{1j}$ and $Y_{2k} > Y_{2j}$ for some $k = 1, \ldots, n\}$ is empty, the sets E_1 and E_2 of Lemma 4 are each non-empty, and $R_0(Y_{1i}, Y_{2j}) > 0$] or [the set $F_1' = \{(x,y) : (y,x) \in F_1\}$ is empty, the sets E_1' and E_2' are each non-empty, and $R_0(Y_{1i}, Y_{2j}) > 0$].*

Proof: This covers plot 16 of Figure 1. Note $R_3(Y_{1i}, Y_{2j}) = 0$, so that negative mass is assigned if and only if $[R_0(Y_{1i}, Y_{2j}) > 0, R_1(Y_{1i}, Y_{2j}) < 0$, and $R_2(Y_{1i}, Y_{2j}) > 0]$ or $[R_0(Y_{1i}, Y_{2j}) > 0, R_1(Y_{1i}, Y_{2j}) > 0$, and $R_2(Y_{1i}, Y_{2j}) < 0]$. By Lemma 2, $R_1(Y_{1i}, Y_{2j}) < 0$ and $R_2(Y_{1i}, Y_{2j}) > 0$ if and only if $F_1 = \emptyset, E_1 \neq \emptyset$, and $E_2 \neq \emptyset$. □

Lemma 6 (Type IV) *Negative mass is not assigned to any points not covered by Lemmas 3–5.*

Proof: This covers plots 3, 7, 9, 10, and 11 of Figure 1 and the case of mass assigned to an uncensored point. First consider the case when $Y_{2i} < Y_{2j}, \delta_{2i} = 0$, and $Y_{1i} < Y_{1j}$ if $\delta_{1j} = 1$. Here $R_3(Y_{1i}, Y_{2j}) = 0, R_2(Y_{1i}, Y_{2j}) > 0, R_1(Y_{1i}, Y_{2j}) > 0$, and $R_0(Y_{1i}, Y_{2j}) \geq 0$. To complete the proof consider the case of mass assigned to an uncensored point. In this case $R_1(Y_{1i}, Y_{2j}) > 0, R_2(Y_{1i}, Y_{2j}) > 0$, and $R_3(Y_{1i}, Y_{2j}) \leq 0$. □

Remark 1: If $\hat{L}(\Delta u, \Delta v) = 1$ — instead of zero — when (1) is not well-defined, Lemmas 2–5 go through unchanged, but negative mass may be assigned to an uncensored point (s,t) if $n_{s,t} = 1$.

Remark 2: Points of type III are only due to edge effects, but points of types I and II will generally be quite common, and can never disappear with the addition of more observations. By examining only points of plots 1, 4, and 13 (of which there will generally be $O(n^2)$), it is heuristically clear that the number of points assigned negative mass increases as n^2.

Remark 3: Negative mass can be redistributed while maintaining Kaplan-Meier marginals in a number of ways, and if the total amount of negative mass is negligible, these *ad hoc* estimates will inherit properties from Dabrowska's estimate. However, from examining computer generated data and heuristic considerations this is not the case. If Y_i and Y_j are uncensored observations, the magnitude of negative mass assigned to the point in plot 1 (or plots 4 and 13) can be bounded below by $R_0(s,t)\hat{F}(s,0)\hat{F}(0,t)n^{-2}(k_1/n_{s,0+})(k_2/n_{0+,t})$ where $s = Y_{11} \wedge Y_{12}, t = Y_{21} \wedge Y_{22}, k_1 = \#\{(Y_{1k}, Y_{2k}) : Y_{1k} > s, Y_{2k} < t$, and $\delta_{2k} = 0$ for some $k = 1, \ldots, n\}$, and $k_2 = \#\{(Y_{1k}, Y_{2k}) : Y_{1k} < s, Y_{2k} > t$, and $\delta_{1k} = 0$ for some $k = 1, \ldots, n\}$. The expected value of this is $O(n^{-2})$.

Reference

DABROWSKA, D.M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* To appear.

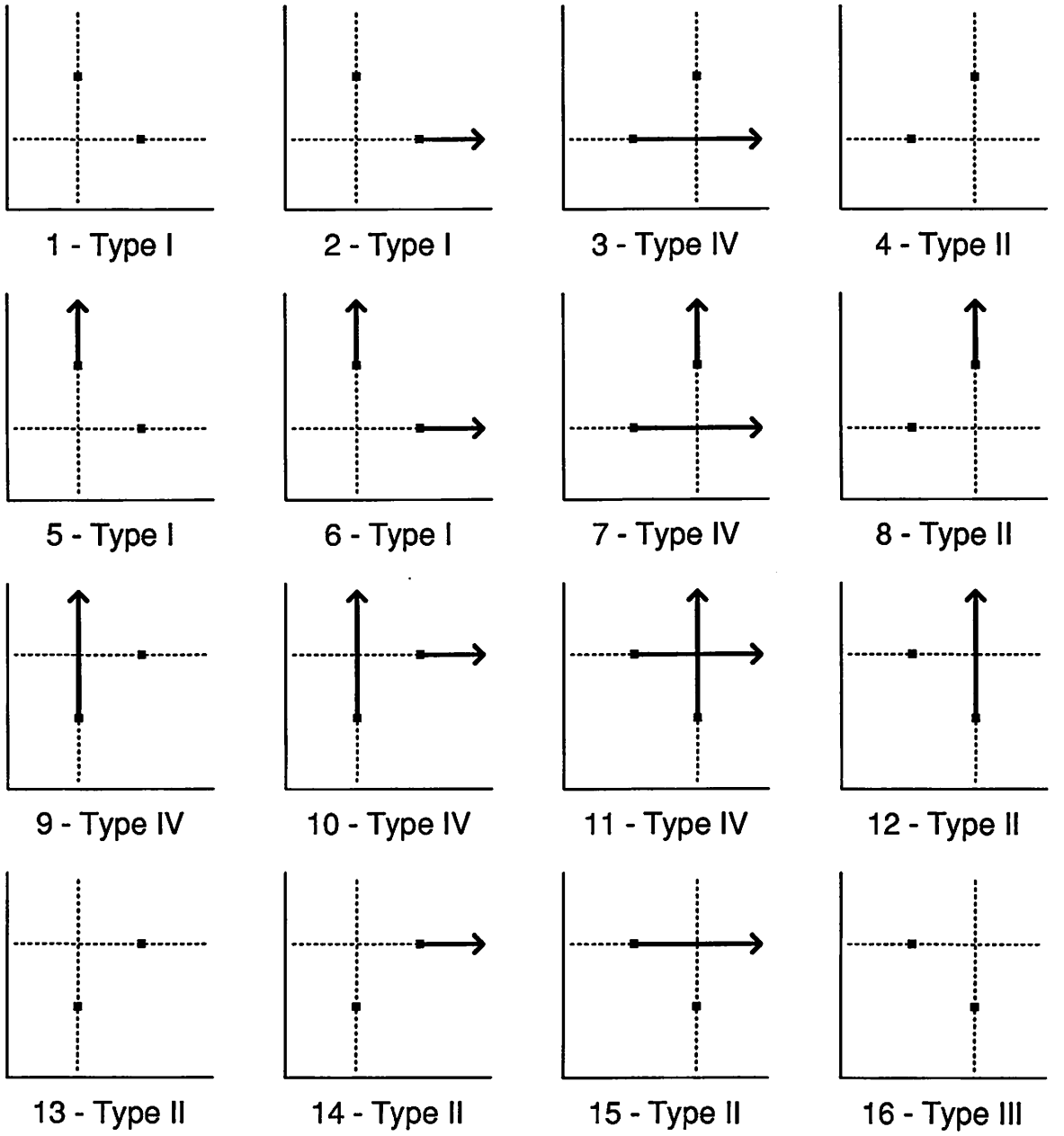

Figure 1. Points -- indicated by the intersection of the dashed lines -- of possible non-zero mass for Dabrowska's estimate.